Eur. J. Clin. Chem. Clin. Biochem.
Vol. 29, 1991, pp. 813–818

The Reproducibility of Urinalysis Using Multiple Reagent Test Strips

By *R. A. G. Winkens*[1], *P. Leffers*[2], *C. P. Degenaar*[3] and *A. W. Houben*[4]

[1] *Diagnostic Centre, Maastricht*
[2] *Department of Epidemiology, State University of Limburg, Maastricht*
[3] *Department of Clinical Chemistry, University Hospital, Maastricht*
[4] *Department of Medical Microbiology, State University of Limburg, Maastricht*

(Received August 6, 1990/September 16, 1991)

Summary: Ninety urine samples were examined twice by 3 "observers" (two persons, using only visual observation, and one person using a spectrophotometric analyser) using multiple reagent teststrips. To determine reproducibility, inter- and intra-observer agreement were calculated and expressed as *Cohen*'s kappa and as weighted kappa.

The results show negligible intra-observer differences between the visual and spectrophotometric observation. The lack of agreement between inter- and intra-observer urinalysis results, using multiple reagent test strips was disappointing, considering the simplicity of the test procedure. Further improvement of reproducibility, e. g. by enhancing the discoloration of the test pads, is necessary. Reproducibility is not improved by using a spectrophotometric analyser instead of visual reading of the test strips.

Introduction

Urinalysis is very frequently performed in clinical chemistry laboratories. The test problems for urinalysis can be divided into two main groups:

- microscopic examination of the urinary sediment
- examination using multiple reagent test strips.

The urinary sediment used to be a widely used diagnostic tool. However, sediment analysis is subject to many sources of error, and it has a rather low reliability (1, 2).

Several years ago a multiple reagent test strip was developed. The discoloration of its test pads after immersion in the urine sample can be visually judged or measured by means of a spectrophotometer. Since it is very easy to use, it appears to be a suitable alternative to the analysis of urinary sediment. When used with the spectrophotometer it is considered to be "an almost ideal test: simple to perform, quick, inexpensive and easy to interpret" (3). The literature however reveals little information about its diagnostic performance (4–8). Therefore we thought that an additional evaluation was required.

In the evaluation of a diagnostic tool both its diagnostic value and its reproducibility should be determined (9). The diagnostic value is the certainty with which a positive or negative test result predicts the presence or absence of a disease. Reproducibility is the extent to which the test leads to the same result when performed by different analysts using the same or different reading techniques (inter-observer agreement) or when performed by the same analyst (intra-observer agreement).

A diagnostic test can only have a high diagnostic value if reproducibility is good. Low reproducibility may mean that the test procedure can still be improved, thereby also improving the diagnostic value of the test.

The reproducibility of urinalysis can be considered to be good, if repeated testing leads to the same results, and it can be assumed that no real change has occurred in the urine sample between the first and last test.

Changes may result from ageing of the sample (dependent on storage temperature, pH and osmolality). Also, alterations of the test strip itself (expiry date, storage temperature) and variation through differences in the degree of homogenization of the urine sample may affect test results (10–12).

Reproducibility is influenced by subjectivity in the grading of the test result and by differences in the execution of the test (13). In the laboratory, changes in the urine sample and in the reagent test strip can be prevented (2, 14, 15). Observational errors are much more difficult to control. Therefore, we aimed our study at the influence of observational errors on test reproducibility. In order to measure inter- and intra-observer variation, we carried out a study, using visual and spectrophotometric reading of multiple-reagent test strips for the examination of selected urine samples.

Overall agreement as a measure of observer variation has the drawback that even if the observers randomly assign test results, there could still be agreement by chance. The level of this chance agreement depends on the prevalence of positive test results in the study population. We used *Cohen*'s kappa (κ), which is a measure of reproducibility corrected for "agreement by chance" (15, 17) (see appendix).

Cohen's kappa treats all disagreement in the same way, independent of the distance of the test result on the ordinal scale. As one might argue that a measure of reproducibility should take account of the distance between the test results, weighted kappa (κ_w) was also calculated (18).

Methods

For the purpose of the experiment, all the urine samples collected through inpatient- and outpatient clinics and delivered daily at the department of clinical chemistry of our university hospital were screened, and 90 samples were selected on the basis of a positive reaction for one or more of the following tests: leukocyte esterase activity, nitrite, blood and protein.

In the experiment, a seven-patch test strip for the determination of leukocyte esterase activity, nitrite, pH, protein, glucose, ketone bodies and blood (Nephur-7-RL, Boehringer Mannheim, Almere, the Netherlands) was used for all measurements. The test strips were used according to the manufacturer's recommendations.

All selected urine samples were examined twice by 3 "observers":

- an experienced laboratory technician (visual observation)
- a non-experienced laboratory-school student (visual observation)
- a spectrophotometric analyser, Urotron RL9 (Boehringer Mannheim, Almere, the Netherlands).

To prevent recognition of urine samples, the sequence of the samples was changed after the first series of measurements, using a list of random numbers. To avoid any influence of "ageing" of the urine sample, the first and second measurement of every urine sample were performed within one hour. All test results were graded as −, +, ++ or +++. We adjusted the cut-off levels of the spectrophotometric analyser, so that they would match those for visual observation.

For the determination of reproducibility, inter- and intra-observer agreement were expressed as *Cohen*'s kappa and weighted kappa (16, 18).

Results

Among the selected 90 urine samples, positive test results were obtained for leukocyte esterase activity in 36 samples (40%), nitrite in 13 samples (14.5%), protein in 22 samples (24.5%) and blood in 39 samples (43.5%).

Inter-observer agreement is shown in table 1. Kappa ranges from 0.34 to 0.98. The highest agreement was achieved for the nitrite reaction, the lowest agreement was achieved for the determination of glucose.

The highest average inter-observer agreement (0.81) was achieved by the two "visual" observers. For both the laboratory technician and the laboratory-school student, agreement with the spectrophotometric analyser was strikingly low, with an average of 0.59 and 0.57 respectively.

Intra-observer agreement is shown in table 2. Kappa varied from 0.57 to 1.0. Again, the highest agreement was achieved for the nitrite reaction, but the lowest agreement was achieved for leukocyte esterase activity.

With a "mean of kappas" of 0.84, the spectrophotometric analyser achieved the highest average agreement. Although only slightly lower, the experienced laboratory technician achieved the lowest average agreement, with a "mean of kappas" of 0.79.

Since test results sometimes varied by more than one category, we also calculated weighted kappa. For inter-observer agreement weighted kappas are shown in table 3 and for intra-observer agreement in table 4. Weighted kappa varied from 0.67 to 0.98 and from 0.66 to 1.0, respectively. In almost all situations, kappa increases after weighting. For intra-observer agreement, expressed as weighted kappa, all three observers achieved almost the same level of agreement. Surprisingly, the spectrophotometric analyser had the lowest agreement, with a mean of weighted kappas of 0.90.

Tab. 1. Inter-observer agreement for the several pairs of observers, expressed as *Cohen*'s kappa for leukocyte esterase activity, nitrite, acidity (pH), protein, glucose, ketone bodies and blood.

Testpad	Observers		
	1 versus 2	1 versus 3	2 versus 3
Leukocyte esterase activity	0.74	0.57	0.70
Nitrite	0.91	0.98	0.95
pH	0.86	0.52	0.54
Protein	0.82	0.53	0.54
Glucose	0.62	0.46	0.34
Ketone bodies	0.92	0.37	0.36
Blood	0.83	0.72	0.58
Mean	0.81	0.59	0.59

1 = laboratory technician
2 = laboratory school student
3 = spectrophotometric analyser

Tab. 2. Intra-observer agreement, expressed as *Cohen*'s kappa for leukocyte esterase activity, nitrite, acidity (pH), protein, glucose, ketone bodies and blood.

Testpad	Observers		
	1	2	3
Leukocyte esterase activity	0.57	0.61	0.77
Nitrite	0.88	1.0	0.91
pH	0.73	0.80	0.67
Protein	0.74	0.64	1.0
Glucose	0.86	0.87	0.94
Ketone bodies	0.88	1.00	0.75
Blood	0.85	0.75	0.86
Mean	0.79	0.81	0.84

1 = laboratory technician
2 = laboratory school student
3 = spectrophotometric analyser

Tab. 3. Inter-observer variation agreement for the several pairs of observers, expressed as weighted kappa for leukocyte esterase activity, nitrite, acidity (pH), protein, glucose, ketone bodies and blood.

Testpad	Observers			Mean
	1 ↔ 2	1 ↔ 3	2 ↔ 3	
Leukocyte esterase activity	0.85	0.88	0.86	0.84
Nitrite	0.91	0.98	0.95	0.95
pH	0.91	0.84	0.68	0.81
Protein	0.94	0.82	0.86	0.87
Glucose	0.93	0.81	0.74	0.83
Ketone bodies	0.98	0.67	0.73	0.80
Blood	0.95	0.93	0.89	0.92
Mean	0.92	0.85	0.82	

1 = laboratory technician
2 = laboratory school student
3 = spectrophotometric analyser

Tab. 4. Intra-observer agreement, expressed as weighted kappa for leukocyte esterase activity, nitrite, acidity (pH), protein, glucose, ketone bodies and blood.

Testpad	Observers			Mean
	1	2	3	
Leukocyte esterase activity	0.85	0.92	0.94	0.90
Nitrite	0.88	1.0	0.91	0.93
pH	0.92	0.88	0.88	0.89
Protein	0.92	0.97	1	0.93
Glucose	0.98	0.94	0.93	0.96
Ketone bodies	0.98	1	0.66	0.88
Blood	0.96	0.94	0.97	0.95
Mean	0.93	0.94	0.90	

1 = laboratory technician
2 = laboratory school student
3 = spectrophotometric analyser

Discussion

We chose kappa and weighted kappa as measures for inter- and intra-observer agreement (16–18). They are now accepted measures in the evaluation of reproducibility in clinical medicine. They express the extent that agreement exceeds the agreement achieved by chance. Although there exists no objective interpretation, kappa under 0.40 is interpreted as low agreement; kappa between 0.40 and 0.75 is interpreted as moderate to reasonable agreement and kappa higher than 0.75 is interpreted as good agreement (16, 17). We believe that agreement should be *good* if a test is to be applied in clinical practice.

The changes that can occur within the urine-sample itself are relatively well known (2, 14, 15). Before the experiment, all urine samples were stored below 8 °C.

During the experiment, all urine samples were homogenized before examination and every sample was examined twice by every observer within one hour. In this period, relevant alterations are very unlikely to occur. Therefore, we can presume that varying test results for each urine sample are not caused by alterations within the urine sample itself.

Visual observation on the other hand, in which the amount of experience and education may play an important role, is less controllable. Overall there is only a negligible difference in intra-observer agreement between the two visual observers and the spectrophotometer. Although it is not possible to draw hard conclusions, the data do not show a positive influence of experience on intra-observer agreement.

For the majority of test-pads, inter-observer agreement is not high enough (lower than 0.75). Adequate agreement between the three observers was achieved only for nitrite. There is, however, a remarkable agreement between the laboratory technician and the laboratory-school student. Their agreement with the spectrophotometric analyser is considerably lower. This striking difference may (at least partly) be explained by the following: Although we adjusted the cut-off levels of the spectrophotometric analyser so that they matched those for visual observation, small differences cannot be ruled out. Obviously, this could only influence inter-observer agreement between the spectrophotometric analyser and both visual observers.

It should be realized that the higher inter-observer agreement between the two persons does not mean that their readings reflect the composition of the urine sample more validly than the readings from the spectrophotometric analyser. It is quite possible that a good agreement is achieved, despite inaccurate observations, when two observers make the same mistake in the same measurement.

Intra-observer agreement permits insight into the performance of each separate "observer". The highest intra-observer agreement could be expected for the spectrophotometric analyser, which is not impeded by factors like lack of experience, tiredness etc.

Nevertheless, intra-observer agreement for the spectrophotometric analyser is not always perfect.

The performance of the two "visual" observers is hardly worse than that of the spectrophotometric analyser (the mean of kappas is 0.79, 0.81 and 0.84 respectively).

Urinalysis by a spectrophotometric analyser results in only a minor improvement of reproducibility. Our data do not confirm the (generally accepted) assumption that "automation" of urinalysis improves the reproducibility of urine examination.

Perhaps large discrepancies in the test strip readings should be penalized more harshly than small ones. Therefore, we also calculated weighted kappa. The result was a considerable increase in kappa-values for both inter- and intra-observer agreement. This points to the fact that in general the disagreements did not exceed one category on the ordinal scale.

However, for clinical practice we believe that even such a small disagreement is also important and should not be tolerated.

In general, the kappa-values we calculated demonstrate that reproducibility of urinalysis with test strips leave room for improvement. How this improvement can be achieved is not yet clear. Our results show that it is not likely to be achieved by using a spectrophotometric analyser. For visual test strip reading it would help if the degree of discoloration of test pads, in particular for leukocyte esterase activity and glucose, could be enhanced.

Conclusions

Reproducibility of urinalysis by using multiple reagent test strips is (as a rule) moderate to good. However, for such a simple test procedure, one should not be satisfied with a reproducibility which is (only) moderate to good. Intra-observer agreement for spectrophotometric analysis is only marginally higher than intra-observer agreement with visual observation. In view of these small differences, application of a spectrophotometric analyser is not a matter of course. Further efforts to enhance reproducibility should be encouraged. At present, the results of urinalysis with test strips is still quite dependent on when, and by whom the reading is performed and on what equipment is used.

Acknowledgement

We thank Boehringer Mannheim b. v., Almere, The Netherlands for providing us with the test strips and Dr. *H. Schouten* for his statistical advice.

Appendix

Suppose two observers perform a test on N cases. The test can give outcomes with k possibilities.

Judging the cases leads to the following table with chances for (dis)agreement.

	Observer A (i)				
Observer B (j)	1	2	. .	k	total
1	P_{11}	p_{12}	. .	p_{1k}	$P_{1.}$
2	P_{21}	P_{22}	. .	P_{2k}	$P_{2.}$
	P_{k1}	P_{k2}	. .	P_{kk}	$P_{k.}$
Total	$P_{.1}$	$P_{.2}$	. .	$P_{.k}$	1

Observed agreement = $p_{11} + P_{22} + \ldots + P_{kk}$
Expected chance agreement = $P_{1.} P_{.1} + P_{2.} P_{.2} + . + P_{k.} P_{.k}$

Kappa corrects for the "agreement by chance" in the following way:

$$\text{Kappa} = \frac{\text{observed agreement (\%)} - \text{expected chance agreement (\%)}}{100\% - \text{expected change agreement}}$$

$$\text{i.e. } \kappa = \frac{P_o - P_e}{1 - P_e}$$

Kappa can vary from -1 up to $+1$.

A negative kappa means that the agreement is less than that expected from chance. A kappa-value of 0 means that the agreement is equal to the expected chance agreement, and kappa larger than 0 means that the agreement is higher than the expected chance agreement. Weighted kappa also takes into account the severity of the disagreement in observations on an ordinal scale.

Since observers do or do not agree with each other, the proportion of disagreement Q can be seen as 100% minus the proportion of agreement P, and therefore $Q = 1 - P$.

The equation for kappa can then be changed into:

$$\frac{(1-Q_o) - (1-Q_e)}{1 - (1-Q_e)} = \frac{Q_e - Q_o}{Q_e}.$$

Then weighted kappa

$$\kappa_w = 1 - \frac{Q_o}{Q_e}.$$

Q_o is calculated by multiplying every disagreement-cell proportion where $i \neq j$ ($P_{12} + \ldots + P_{k-1k} + P_{21} + \ldots + P_{kk-1}$) with a weight factor V_{ij} and summing the products.

Q_e is calculated by summing the products of the proportions for the row and column of every disagreement cell ($P_{1.} P_{.2} + \ldots + P_{1.} P_{.k + \ldots + Pk.} P_{.1} + \ldots + P_{k.} P_{.k-1}$) and multiplying them with a corresponding weight factor V_{ij}.

When the results of two observations are compared in a cross-table, each cell receives a weight factor (normally the difference raised to the square): $V_{ij} = (i-j)^2$.

Weighted kappa is then calculated by the equation:

$$\kappa_w - \frac{\Sigma V_{ij} P_{oij}}{\Sigma V_{ij} P_{eij}}$$

In this equation, V_{ij} is the disagreement weight, P_{oij} the observed cell proportion for disagreement and P_{eij} is the expected chance cell proportion for disagreement.

References

1. Winkel, P., Statland, B. E. & Jørgensen, K. (1974) Urine Microscopy, an Ill-defined Method, Examined by a Multifactorial Technique. Clin. Chem. *20*, 436–439.
2. Gadeholt, H. (1964) Quantitative Estimation of Urinary Sediment, with Special Regard to Sources of Error. Br. Med. J. *1*, 1547–1549.
3. Fraser, C. G. (1985) Urine Analysis: Current Performance and Strategies for Improvement. Br. Med. J. *291*, 321–323.
4. Yamane, N., Sakamoto, F. & Matsuura, F. (1988) Quantification of Urinary Glucose and Protein with Test-strips through Reflectometric Analysis. Clin. Biochem. *21*, 271–275.
5. James, G. P., Bee, D. E. & Fuller, J. B. (1978) Accuracy and Precision of Urinary pH Determinations using Two Commercially Available Dipsticks. Am. J. Clin. Pathol. *70*, 368–374.
6. Gupta, R. C., Goyal, A. & Singh, P. P. (1982) Reliability of urinalysis for glucose. Clin. Chem. *28*, 1724 (letter).
7. Simpson, E. & Thompson, D. (1978) An assessment of hospital routine urinalysis. Ann. Clin. Biochem. *15*, 241–242.
8. Marx, A. M., Kropf, J. & Gressner, A. M. (1989) On the performance and reliability of mechanized urine teststrip measurement in comparison with visual reading. J. Clin. Chem. Clin. Biochem. *27*, 433–443.
9. Sackett, D. L., Haynes, R. B. & Tugwell, P. (1985) *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston: Little, Brown and Company.
10. Triger, D. R. & Smith, J. W. G. (1966) Survival of Urinary Leucocytes. J. Clin. Pathol. *19*, 443–447.
11. Vaughn, E. D. & Wyker, A. W. (1971) Effect of Osmolality on the Evaluation of Microscopic Hematuria. J. Urol. *105*, 709–711.

12. Nanji, A. A., Poon, R. & Hinberg, J. (1988) Effect of not Allowing Reflotron Strips to Warm to Room Temperature (techn. brief). Clin. Chem. *34*, 179–180.
13. Spodick, D. H. (1975) On Experts and Expertise: the Effects of Variability in Observer Performance. Am. J. Cardiol. *36*, 592–596.
14. Kierkegaard, H., Feldt-Rasmussen, U., Horder, M., Andersen, H. J. & Jørgensen, P. J. (1980) Falsely Negative Urinary Leucocyte Counts Due to Delayed Examination. Scand. J. Clin. Lab. Invest. *40*, 259–261.
15. Hindman, R., Tronic, B. & Bartlett, R. (1976) Effect of Delay on Culture of Urine. J. Clin. Microbiol. *4*, 102–103.
16. Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement *20*, 37–46.
17. Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. New York: Wiley & Sons.
18. Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bulletin *70*, 213–220.

R. A. G. Winkens
Diagnostic Centre Maastricht
P. O. Box 1918
NL-6201 BX Maastricht